This Issue Completes Volume 43
Volume 43, Issue 6, December 2009

ISSN 0092-6566

ELSEVIER

JOURNAL OF
RESEARCH IN
PERSONALITY

Editor-in-Chief
RICHARD E. LUCAS

Associate Editors
PETER BORKENAU PHEBE CRAMER
M. BRENT DONNELLAN

Available online at
ScienceDirect
www.sciencedirect.com

Journal of Research in Personality 43 (2009) 1053–1063

Contents lists available at ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

Evaluating correlations in studies of personality and behavior: Beyond the number of significant findings to be expected by chance

Ryne A. Sherman *, David C. Funder

Department of Psychology, University of California, 900 University Ave., Riverside, CA 92521, USA

ARTICLE INFO

Article history:
Available online 2 June 2009

Keywords:
Randomization tests
Data analysis
Family wise error rate
Statistical significance

ABSTRACT

When large numbers of statistical tests are computed, such as in broad investigations of personality and behavior, the number of significant findings required before the total can be confidently considered beyond chance is typically unknown. Employing modern software, specially written code, and new procedures, the present article uses three sets of personality data to demonstrate how approximate randomization tests can evaluate (a) the number of significant correlations between a single variable and a large number of other variables, (b) the number of significant correlations between two large sets of variables, and (c) the average size of a large number of effects. Randomization tests can free researchers to fully explore large data sets and potentially have even wider applicability.

1. Introduction

When large numbers of statistical tests are computed, particularly in large data sets, some will attain significance by chance alone. Take, for example, a researcher who is interested in determining how a particular behavior relates to personality. The behavioral score – such as a count of how many times an individual used "certainty" words in an hour of speech – may be correlated with ratings of as many as 100 different personality attributes.[1] Some of these correlations – sometimes many – will turn out to be statistically significant, and may be reported in a table. Such a table is often highly interesting, but it can raise concerns with reviewers, readers, and even the researchers themselves, who must all ponder two questions. (1) How many items would appear on this table of significant correlates by chance alone? (2) How many *more* significant correlations than this number are required to justify confidence that the findings, as a set, are non-random? In common research practice the answer to the first question is only approximately estimated and the second is completely unknown.

The same issues arise in any exploratory study in which a behavior is correlated to a large number of personality items, a personality item is correlated to a large number of behaviors, or – more generally – any large number of variables of one kind is compared to variables of another kind. Such studies provide the kind of richly descriptive data that personality psychology badly needs (Funder, in press), but can be difficult to evaluate.

As an example, an article recently published in the *Journal of Personality and Social Psychology* (Fast & Funder, 2008) — co-written by an author of the present paper — concluded that use of certainty words (e.g. absolutely, exact, guarantee, sure, etc.) and use of sexuality words (e.g. boobs, butt, kiss, horny, etc.) in a life-history interview were related to personality characteristics of the speakers as well as to their directly-observed behavior in a separate context. The conclusion that personality was related to word use was based on the evidence that 26 (out of 100) self-reported personality traits were, at the .05 alpha level, statistically significantly related to the use of certainty words and 16 (out of 100) self-reported personality traits were statistically significantly related to the use of sexuality words. Likewise, the conclusion that behavior was related to word use was based on the evidence that 31 (out of 64) observed behaviors were statistically significantly related to the use of certainty words and 23 (out of 64) observed behaviors were statistically significantly related to the use of sexuality words. While the sheer number of significant correlates appears impressive, its statistical interpretation is unclear and can raise – and has raised – concerns from reviewers and others that the findings might capitalize on chance.

The solution to this problem is not to stop conducting broad, descriptive studies. In fact, such research is often necessary (see Block, 1960; Funder, in press). No doubt, the general question "How does behavior relate to personality?" is important. But how many statistically significant relations should the researcher expect by chance? And more importantly, at what point should the researcher feel confident that the set of relationships between a variable of interest and another set of variables are unlikely to have arisen by chance?

These two questions are important to researchers, reviewers, and readers of the scientific literature. From the researcher's

* Corresponding author.
E-mail address: rsher002@ucr.edu (R.A. Sherman).

[1] 100 happens to be the number of personality-descriptive items in the California Adult Q-set (Block, 1961; 1978; 2008), a widely-used research tool.

0092-6566/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jrp.2009.05.010

perspective, it would be poor science to fail to investigate the many possible relationships between personality and behavior. But from the perspective of both the researcher and the reviewer, publishing results which reflect capitalization on chance would be equally poor science. Additionally, a reader of such results, once published, may wish to know the probability of such findings as part of his or her evaluation of their scientific importance. If readers and researchers become confident that the relationships between a variable of interest and a set of other variables go far beyond chance, they may invest time, effort, and resources in pursuing these relationships in further studies; otherwise they may turn to other lines of research.

The general problem, known as the multiple testing or family wise error rate (FWE) problem, was first identified long ago (Cournot, 1843), and entire books have been written on the topic (e.g. Miller, 1966; Westfall & Young, 1993). However, nearly all of the literature has focused on adjusting *p*-values for a single hypothesis test rather than estimating the overall probability of a set of results. In a simple case, if a researcher conducts two *t*-tests without an *a priori* hypothesis, he or she might halve the critical *p*-value (e.g., from .05 to .025). This simple correction, known as the Bonferroni or ensemble adjusted correction (see Rosenthal & Rubin, 1984), while useful in the aforementioned situation, is not very useful to researchers who are interested in evaluating a large set of significant results. For example, Bonferroni adjustment of the evaluation of 100 personality correlations would require imposing a critical *p*-value on each correlation of .0005 – an almost insurmountable threshold. Such an adjustment also misses the main point. Bonferroni and most work on FWE has concerned itself with the question, "Which of the multiple tests should be considered statistically significant?" while the question often of interest when evaluating widely descriptive studies is, "Is there substantial evidence that a series of significant results goes beyond what could be expected by chance?"

By and large, two methods have been employed to address the latter question. The first, method is to (implicitly) assume independence in the series of tests and to determine the probability of finding n results statistically significant given *n* tests using the binomial expansion (Brozek & Tiede, 1952; Sakoda, Cohen, & Beall, 1954; Wilkinson, 1951). For example, out of 100 correlations one might simply expect 10 to be significant at the .10 level, five at the .05 level, and 1 at the .01 level. But Block (1960) criticized the binomial expansion model on the grounds that the independence assumption is rarely, if ever, met.

> ...for many kinds of data the crucial assumption of independence of outcomes of the statistical tests is in serious error because the variables or events providing the data for the statistical tests tend to be correlated (p. 370).

Moreover, even if the number of statistically significant results expected by chance were to be accurately calculated, answering the first question with which this article began, the second question would still remain unanswered: how many statistically significant findings *beyond* this chance level need be identified to have confidence in the results? The answer to this more pressing question lies in identifying a sampling distribution of possible results for a given study. The approximate randomization test, described herein, provides an answer to this question.

A randomization test is a resampling procedure[2] in which a distribution is empirically generated by permuting the original data into other possible data sets. Fisher (1935) demonstrated an approach of this sort in the now classic, "Lady Tasting Tea" example where the task is to determine which 4 of 8 total cups of tea had milk added first. As Fisher described, there are exactly 70 possible ways to classify 8 cups of tea into 2 groups of 4 each. To determine the exact probability, or *p*-value, of a particular number of correct classifications one divides the number of possible ways to obtain the observed result (or a more rare result) by the total number of possible permutations. For example, there is only one possible way to identify all 8 cups correctly and so the probability, or *p*-value, for this result is 1/70 or approximately .014.

Fisher (1935) believed that, in order to be of any use, *p*-values derived from the theoretical distributions employed by parametric statistics (e.g. *t*-tests, *F*-tests) ought to mirror those values derived from empirical distributions. However, he and others since (Edgington, 1969) recognized that while an exact permutation test was appropriate for small sample sizes such as the Lady Tasting Tea example, as sample sizes grow the number of possible combinations of results becomes exceedingly large and the formulation of the exact sampling distribution becomes impossible. Instead, psychologists and statisticians alike have relied on theoretical distributions and parametric statistics. This is practice is valid so long as the empirical data meet the theoretical assumptions of the statistical models – which they often do, but sometimes do not. However, in the age of the high speed computer, it is no longer necessary to rely solely on parametric methods. And, in the present case, no parametric statistics exist to evaluate entire sets of results, such as numerous correlations between attributes of personality and a particular behavior.

To remedy this situation, Block (1960) proposed using an approximate randomization procedure,[3] via computer simulation, to generate an approximate sampling distribution. A generalization of the randomization test proposed by Block is as follows:

(1) Correlate a given single variable *X* with each of a set of given variables *Y*.
(2) Record the number of correlations between the *X* variable and the *Y* variables that attained statistical significance.
(3) Randomly re-assign (without replacement) the participants' scores on *X* to the original set of *Y* scores, creating a new pseudo-sample.
(4) Correlate the randomly assigned *X* variable with each of the original Y variables.
(5) Record the number of significant relationships between *X* and the set of *Y* variables in the pseudo-sample.
(6) Repeat Steps 3 through 5 a large number of times (perhaps 1000 to 10,000). This procedure produces an approximate chance sampling distribution of significant results.
(7) Compare the number of significant results observed in Step 2 to the sampling distribution produced in Step 6 to evaluate the probability of obtaining the observed number of significant findings.

Block (1960) demonstrated this technique in 8 samples using dichotomous response data from true–false inventories. His results suggested that conclusions prescribed by the binomial expansion method might be overly conservative. That is, the average number of statistically significant results one may expect to find by chance when conducting a large number of inference tests did not appear to be as high as the number indicated by the binomial expansion.

[2] For a taxonomy of resampling procedures, see Rodgers (1999).

[3] The terms randomization and permutation are used almost interchangeably throughout the literature, although there is one apparent difference. A permutation test creates pseudo-samples or data sets to form a sampling distribution without replacement such that no pseudo-sample is allowed to appear twice in the sampling distribution whereas a randomization test samples with replacement such that pseudo-samples may appear more than once. When random iterations are run many times, the empirical difference is almost unnoticeable.

Thus, Block's article was a first attempt to address the first question of this article: How many statistically significant results should be expected by chance?

Despite this powerful demonstration, and while there is no doubt that the published literature is still filled with studies correlating many variables with each other, studies which employ the randomization test suggested by Block remain exceedingly rare. In fact, as of this writing, the Social Science Citation Index indicates that Block's (1960) article has been cited 33 times and only 11 of those articles actually employ the method he described, most of which were published in the 1960s, with one notable exception (Fast & Funder, in press). Several other articles cite Block's article but rather than use his randomization method, the nominal level (i.e. if computing 100 tests, 5 expected by chance at $\alpha = .05$) is referred to, an apparent regression to the very binomial expansion technique that Block's article questioned.

The randomization test outlined by Block would seem to apply to a wide range of research, so why has it been so seldom used? One explanation might be Block's admission regarding the randomization test that, "It is perhaps still too early for a strong generalization" (p. 377). Another possibility is that Block's original analyses of dichotomous data have not yet been extended to continuous data. But we believe the main reason that Block's paper failed to have widespread impact is that it was too far ahead of its time. The computer Block used for his analysis was primitive by modern standards, and the programming required was arduous and not generalizable across platforms.[4] As high speed computers with power unimaginable in 1960 began to appear on the desks (or even in the pockets) of nearly all researchers, randomization tests to take advantage of this power were not taught, nor was the necessary software developed to make such analyses easy to conduct.

The present article has several purposes. The first is to highlight the randomization test and its utility for researchers using and readers evaluating research that simultaneously analyses large numbers of variables. In doing so, we provide reproducible programming code in a widely available statistical package (R, see Appendix A). A second purpose is to extend the randomization method described by Block to continuous data and to assess the generalizability of his suggestion that estimating the number of significant findings to be expected by chance by the nominal level may be conservative. Moving beyond Block's contribution, we confront the even more pressing and previously un-addressed concern of determining how many statistically significant findings, beyond this chance expectation, need be identified to have confidence in a series of results. The third purpose of this article is to extend the method to more complex research designs employing multiple predictor variables as well as multiple outcomes, and to outline a procedure that can be feasibly employed in such studies. Finally, this article extends the method to focus on effect sizes rather than arbitrary significance levels.

2. Data set

The data for the following examples come from the Riverside Accuracy Project Phase I (RAP-I). The data include self-reported personality surveys from over 170 undergraduate students at the University of California, Riverside gathered over a 3 year period from October of 1990 until June of 1993. Additionally the data include ratings of the directly-observed behavior of 167 of these participants in three different 5 min interactions with a previously unacquainted person of the opposite sex. The extensive nature of the data set makes it suitable for widely descriptive research and for the examples presented here. Data from the RAP-I have previously been published (Blackman & Funder, 1996; Creed & Funder, 1998a,b; Eaton & Funder, 2001; Eaton & Funder, 2003; Funder, 1995; Funder, Kolar, & Blackman, 1995; Furr & Funder, 1998, 2004; Markey, Funder, & Ozer, 2003; Schimmack, Oishi, Furr, & Funder, 2004; Spain, Eaton, & Funder, 2000); however, the present analyses are entirely new and are used here only to demonstrate the randomization procedures.

The measures relevant to the present analyses are self reports of the Big Five measured by the NEO-PI (Costa & McCrae, 1985), self ratings of personality as measured by the California Adult Q-Sort (CAQ: Block, 1961, 1978, 2008), and behavior as directly-observed from an interaction with an opposite sex stranger in a 5 min unstructured interaction measured by the Riverside Behavioral Q-Sort version 2 (RBQ: Funder, Furr, & Colvin, 2000).[5] The NEO-PI is a widely used measure of the Big Five personality traits. The CAQ consists of 100 wide-ranging personality characteristics (e.g. "Is concerned with philosophical problems", "Is basically anxious"). Participants in the RAP-I rated themselves on these 100 items using a Q-sort procedure in which items are placed into a forced choice quasi-normal distribution ranging from 1 (*extremely uncharacteristic*) to 9 (*extremely characteristic*). Finally, participants in the RAP-I were scheduled to appear in previously unacquainted mixed sex dyads. When the second participant arrived for the interaction the experimenter switched on a videotape recorder and camera in plain sight, told the participants to "talk about whatever you'd like," and left for 5 min. At the end of 5 min the experimenter returned to the room and stopped the recording.

Upon completion of data collection, the 5 min video tapes were coded for behavior using the RBQ. The RBQ (version 2: Funder et al., 2000) contains 64 items which describe behavior at a mid-level of analysis (e.g. acts playful, smiles frequently, expresses insecurity). Sets of four research assistants coded each participant in the interaction using the Q-Sort technique (described above) to sort the 64 items into a forced choice quasi-normal distribution ranging from 1 (*extremely uncharacteristic*) to 9 (*extremely characteristic*).

3. Example #1: Is a variable of interest related to a set of other variables?

Researchers with a large data set may be interested in knowing whether a particular variable of interest is related to a set of other variables.[6] Block's (1960) original work used the randomization method to answer this question using dichotomous predictor and outcome variables. The method is demonstrated here using continuous data.

Using the RAP-I data set, one possible research question is whether or not each of the Big Five is related to behavior in a 5 min unstructured interaction with an opposite sex stranger. To answer this question, each of the 64 coded behaviors from the RBQ were correlated with self-reported Big Five scores from the NEO-PI ($N = 159$). The resulting correlations are displayed in Tables 1–5 for each of the Big Five respectively. For Extraversion, 24 of these 64 correlations were statistically significant at the .05 alpha level (two-tailed), 1 was significant for Neuroticism, 11 for Open-

[4] In fact, Block (personal communication, 2008) informed us that his analyses were programmed in machine language.

[5] An updated, and more widely applicable version (version 3), of the RBQ is currently available at www.rap.ucr.edu.

[6] The terms dependent and independent variable are intentionally avoided here because the variable of interest and the set of other variables can take the form of either the dependent or independent variable depending on the researcher's interests. For example, a health psychologist with a large number of predictor variables might wish to know if the set of predictor (independent) variables has any relationship to longevity (dependent variable). Or a personality psychologist with a special interest in a particular personality trait (independent variable) may wish to know if the said trait has any relationship to a set of outcomes (dependent variables).

ness, 1 for Conscientiousness, and 12 for Agreeableness. Thus it seems fairly obvious that Extraversion has some relationship to the behavior in the interaction, while Neuroticism and Conscientiousness have very little relevance. However, it is somewhat less clear whether Openness and Agreeableness have an impact. Nominally, slightly more than 3 significant correlations in each set would be expected by chance (via the binomial expansion, .05 × 64 = 3.2), but this calculation requires an assumption of independence that, as Block noted, is doubtful at best.

Table 1
Behavioral Correlates of Extraversion from a 5 min Unstructured interaction.

No. RBQ Item	*N* = 159
01 Aware of being on camera	−.08
02 Interviews Partner(s)	.19*
03 Volunteers Information about Self	.15+
04 Interested in what Partner(s) say	.09
05 Tries to control interaction	.04
06 Dominates interaction	.14+
07 Appears relaxed and comfortable	.15+
08 Exhibits social skills	.34***
09 Reserved and unexpressive	−.33***
10 Laughs frequently	.12
11 Smiles frequently	.11
12 Physically animated; Moves a lot	.16*
13 Seems to like partner(s)	.15+
14 Exhibits awkward interpersonal style	−.28***
15 Compares self to other(s)	−.10
16 High enthusiasm and energy level	.27***
17 Displays wide range of interests	.00
18 Talks at partner(s)	−.15+
19 Expresses agreement frequently	−.08
20 Expresses criticism	−.33***
21 Is talkative	.25**
22 Expresses insecurity	−.26***
23 Physical signs of tension/anxiety	−.27***
24 Exhibits high degree of intelligence	−.15+
25 Expresses sympathy towards partner(s)	.02
26 Initiates humor	.24**
27 Seeks reassurance from partner(s)	.12
28 Exhibits condescending behavior	−.13+
29 Seems likeable	.18*
30 Seeks advice from partner(s)	−.01
31 Appears to regard self as phys. attractive	.17*
32 Acts irritated	−.20*
33 Expresses warmth	.02
34 Tries to undermine/sabotage	.02
35 Expresses hostility	−.09
36 Unusual or unconventional appearance	−.06
37 Behaves in fearful or timid manner	−.26***
38 Expressive in voice, face, or gestures	.26***
39 Interest in fantasy or daydreams	−.05
40 Expresses guilt	−.17*
41 Keeps partner(s) at a distance	−.26***
42 Interest in intellectual/cognitive matters	−.14+
43 Seems to enjoy interaction	.18*
44 Says/does interesting things	.03
45 Says negative things about self	−.08
46 Displays ambition	−.14+
47 Blames others	−.12
48 Expresses self-pity or victimization	−.07
49 Expresses sexual interest	.14+
50 Behaves in cheerful manner	.21**
51 Gives up when faced w/obstacles	−.07
52 Behaves in stereotypical gender style or manner	.02
53 Offers advice	.02
54 Speaks fluently; Expresses ideas well	.20**
55 Emphasizes accomplishments	−.08
56 Competes with partner(s)	−.06
57 Speaks in a loud voice	.06
58 Speaks sarcastically	−.11
59 Makes/approaches physical contact	.09
60 Engages in constant eye contact	.16*
61 Seems detached from interaction	−.28***
62 Speaks quickly	.01
63 Acts playful	.16*
64 Partner(s) seek advice from subject	.03

Note. RBQ Item content is abbreviated.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Table 2
Behavioral correlates of neuroticism from a 5 min unstructured interaction.

No. RBQ item	*N* = 159
01 Aware of being on camera	.07
02 Interviews Partner(s)	.04
03 Volunteers Information about Self	−.01
04 Interested in what Partner(s) say	.01
05 Tries to control interaction	.06
06 Dominates interaction	.05
07 Appears relaxed and comfortable	−.03
08 Exhibits social skills	−.09
09 Reserved and unexpressive	.08
10 Laughs frequently	−.01
11 Smiles frequently	−.00
12 Physically animated; Moves a lot	.05
13 Seems to like partner(s)	−.01
14 Exhibits awkward interpersonal style	.14+
15 Compares self to other(s)	.11
16 High enthusiasm and energy level	−.03
17 Displays wide range of interests	−.06
18 Talks at partner(s)	.02
19 Expresses agreement frequently	−.08
20 Expresses criticism	.06
21 Is talkative	−.08
22 Expresses insecurity	.00
23 Physical signs of tension/anxiety	−.00
24 Exhibits high degree of intelligence	−.02
25 Expresses sympathy towards partner(s)	−.06
26 Initiates humor	−.03
27 Seeks reassurance from partner(s)	−.00
28 Exhibits condescending behavior	−.08
29 Seems likeable	−.07
30 Seeks advice from partner(s)	−.03
31 Appears to regard self as phys. attractive	−.05
32 Acts irritated	.10
33 Expresses warmth	.05
34 Tries to undermine/sabotage	.03
35 Expresses hostility	−.04
36 Unusual or unconventional appearance	.11
37 Behaves in fearful or timid manner	.00
38 Expressive in voice, face, or gestures	−.02
39 Interest in fantasy or daydreams	−.11
40 Expresses guilt	−.09
41 Keeps partner(s) at a distance	.08
42 Interest in intellectual/cognitive matters	.00
43 Seems to enjoy interaction	−.06
44 Says/does interesting things	.03
45 Says negative things about self	.10
46 Displays ambition	.07
47 Blames others	.03
48 Expresses self-pity or victimization	.08
49 Expresses sexual interest	−.12
50 Behaves in cheerful manner	−.04
51 Gives up when faced w/obstacles	.13
52 Behaves in stereotypical gender style or manner	−.12
53 Offers advice	.13
54 Speaks fluently; Expresses ideas well	−.12
55 Emphasizes accomplishments	−.13+
56 Competes with partner(s)	−.04
57 Speaks in a loud voice	−.07
58 Speaks sarcastically	.15+
59 Makes/approaches physical contact	−.14+
60 Engages in constant eye contact	−.18*
61 Seems detached from interaction	.02
62 Speaks quickly	−.10
63 Acts playful	.01
64 Partner(s) seek advice from subject	.02

Note. RBQ item content is abbreviated.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Table 3
Behavioral correlates of openness from a 5 min unstructured interaction.

No. RBQ Item	*N* = 159
01 Aware of being on camera	.03
02 Interviews Partner(s)	.10
03 Volunteers Information about Self	–.06
04 Interested in what Partner(s) say	.02
05 Tries to control interaction	–.02
06 Dominates interaction	.13+
07 Appears relaxed and comfortable	.16*
08 Exhibits social skills	.15+
09 Reserved and unexpressive	.12
10 Laughs frequently	–.20*
11 Smiles frequently	.07
12 Physically animated; moves a lot	.09
13 Seems to like partner(s)	–.01
14 Exhibits awkward interpersonal style	–.12
15 Compares self to other(s)	–.06
16 High enthusiasm and energy level	.16*
17 Displays wide range of interests	.02
18 Talks at partner(s)	–.02
19 Expresses agreement frequently	–.08
20 Expresses criticism	–.19*
21 Is talkative	.15+
22 Expresses insecurity	–.17*
23 Physical signs of tension/anxiety	–.29***
24 Exhibits high degree of intelligence	.07
25 Expresses sympathy towards partner(s)	–.05
26 Initiates humor	.19*
27 Seeks reassurance from partner(s)	–.11
28 Exhibits condescending behavior	.02
29 Seems likeable	.05
30 Seeks advice from partner(s)	–.14+
31 Appears to regard self as phys. attractive	–.04
32 Acts irritated	–.13
33 Expresses warmth	.00
34 Tries to undermine/sabotage	.11
35 Expresses hostility	–.03
36 Unusual or unconventional appearance	–.01
37 Behaves in fearful or timid manner	–.19*
38 Expressive in voice, face, or gestures	.18*
39 Interest in fantasy or daydreams	–.09
40 Expresses guilt	–.06
41 Keeps partner(s) at a distance	–.11
42 Interest in intellectual/cognitive matters	.07
43 Seems to enjoy interaction	.14+
44 Says/does interesting things	.12
45 Says negative things about self	–.13
46 Displays ambition	–.02
47 Blames others	.01
48 Expresses self-pity or victimization	–.03
49 Expresses sexual interest	–.04
50 Behaves in cheerful manner	.10
51 Gives up when faced w/obstacles	–.04
52 Behaves in stereotypical gender style or manner	–.18*
53 Offers advice	–.04
54 Speaks fluently; expresses ideas well	.16*
55 Emphasizes accomplishments	.03
56 Competes with partner(s)	.09
57 Speaks in a loud voice	.12
58 Speaks sarcastically	–.04
59 Makes/approaches physical contact	–.02
60 Engages in constant eye contact	.12
61 Seems detached from interaction	–.11
62 Speaks quickly	–.02
63 Acts playful	–.03
64 Partner(s) seek advice from subject	–.04

Note. RBQ item content is abbreviated.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Table 4
Behavioral correlates of conscientiousness from a 5 min unstructured interaction.

No. RBQ item	*N* = 159
01 Aware of being on camera	.10
02 Interviews Partner(s)	.02
03 Volunteers Information about Self	–.00
04 Interested in what Partner(s) say	.05
05 Tries to control interaction	.03
06 Dominates interaction	.10
07 Appears relaxed and comfortable	–.02
08 Exhibits social skills	.06
09 Reserved and unexpressive	–.11
10 Laughs frequently	.01
11 Smiles frequently	.06
12 Physically animated; moves a lot	.08
13 Seems to like partner(s)	–.02
14 Exhibits awkward interpersonal style	–.04
15 Compares self to other(s)	.01
16 High enthusiasm and energy level	.13
17 Displays wide range of interests	.03
18 Talks at partner(s)	–.07
19 Expresses agreement frequently	–.05
20 Expresses criticism	–.23**
21 Is talkative	.08
22 Expresses insecurity	.09
23 Physical signs of tension/anxiety	–.02
24 Exhibits high degree of intelligence	–.06
25 Expresses sympathy towards partner(s)	–.04
26 Initiates humor	.05
27 Seeks reassurance from partner(s)	–.04
28 Exhibits condescending behavior	–.00
29 Seems likeable	–.02
30 Seeks advice from partner(s)	.11
31 Appears to regard self as phys. attractive	.06
32 Acts irritated	–.07
33 Expresses warmth	.02
34 Tries to undermine/sabotage	.07
35 Expresses hostility	–.09
36 Unusual or unconventional appearance	–.00
37 Behaves in fearful or timid manner	–.02
38 Expressive in voice, face, or gestures	.02
39 Interest in fantasy or daydreams	–.01
40 Expresses guilt	.03
41 Keeps partner(s) at a distance	–.05
42 Interest in intellectual/cognitive matters	–.00
43 Seems to enjoy interaction	–.01
44 Says/does interesting things	–.08
45 Says negative things about self	–.03
46 Displays ambition	.02
47 Blames others	–.12
48 Expresses self-pity or victimization	–.08
49 Expresses sexual interest	–.04
50 Behaves in cheerful manner	.06
51 Gives up when faced w/obstacles	–.13+
52 Behaves in stereotypical gender style or manner	.04
53 Offers advice	.06
54 Speaks fluently; expresses ideas well	–.03
55 Emphasizes accomplishments	.04
56 Competes with partner(s)	.13
57 Speaks in a loud voice	.05
58 Speaks sarcastically	–.06
59 Makes/approaches physical contact	–.07
60 Engages in constant eye contact	.02
61 Seems detached from interaction	–.10
62 Speaks quickly	–.01
63 Acts playful	–.03
64 Partner(s) seek advice from subject	.07

Note. RBQ item content is abbreviated.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Instead, we used the randomization method to assess the probability of finding the obtained number of significant correlations, in each table, by chance. Beginning first with Extraversion, pseudosamples were created by randomly redistributing the original extraversion scores provided by the participants to the behavior profiles without replacement such that each behavior profile had an equal probability of being assigned any one of the 159 extraversion scores and each original score is represented in the original

Table 5
Behavioral correlates of agreeableness from a 5 min unstructured interaction.

No. RBQ Item	$N = 159$
01 Aware of being on camera	−.18*
02 Interviews Partner(s)	.06
03 Volunteers Information about Self	−.01
04 Interested in what Partner(s) say	.12
05 Tries to control interaction	−.10
06 Dominates interaction	.01
07 Appears relaxed and comfortable	.02
08 Exhibits social skills	.21**
09 Reserved and unexpressive	−.17*
10 Laughs frequently	.04
11 Smiles frequently	.07
12 Physically animated; moves a lot	.04
13 Seems to like partner(s)	.04
14 Exhibits awkward interpersonal style	−.12
15 Compares self to other(s)	−.03
16 High enthusiasm and energy level	.13+
17 Displays wide range of interests	−.05
18 Talks at partner(s)	−.02
19 Expresses agreement frequently	.03
20 Expresses criticism	−.25**
21 Is talkative	.05
22 Expresses insecurity	−.15+
23 Physical signs of tension/anxiety	−.13
24 Exhibits high degree of intelligence	.03
25 Expresses sympathy towards partner(s)	.19*
26 Initiates humor	.16*
27 Seeks reassurance from partner(s)	−.01
28 Exhibits condescending behavior	−.24**
29 Seems likeable	.18*
30 Seeks advice from partner(s)	−.06
31 Appears to regard self as phys. attractive	.08
32 Acts irritated	−.16*
33 Expresses warmth	.15+
34 Tries to undermine/sabotage	.03
35 Expresses hostility	−.21**
36 Unusual or unconventional appearance	−.14+
37 Behaves in fearful or timid manner	−.04
38 Expressive in voice, face, or gestures	.15+
39 Interest in fantasy or daydreams	−.07
40 Expresses guilt	.02
41 Keeps partner(s) at a distance	−.13
42 Interest in intellectual/cognitive matters	.07
43 Seems to enjoy interaction	.11
44 Says/does interesting things	−.04
45 Says negative things about self	−.11
46 Displays ambition	.01
47 Blames others	−.00
48 Expresses self-pity or victimization	.08
49 Expresses sexual interest	.07
50 Behaves in cheerful manner	.22**
51 Gives up when faced w/obstacles	.07
52 Behaves in stereotypical gender style or manner	.04
53 Offers advice	−.04
54 Speaks fluently; expresses ideas well	.15+
55 Emphasizes accomplishments	.03
56 Competes with partner(s)	−.04
57 Speaks in a loud voice	.08
58 Speaks sarcastically	−.15+
59 Makes/approaches physical contact	.02
60 Engages in constant eye contact	.20*
61 Seems detached from interaction	−.10
62 Speaks quickly	−.03
63 Acts playful	.06
64 Partner(s) seek advice from subject	−.04

Note. RBQ item content is abbreviated.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

data set. The extraversion scores in these pseudo-samples were then correlated with each of the 64 behaviors and the number statistically significant at the .05 level was recorded. The procedure of random re-assignment, correlation computation, and recording the number significant was repeated 10,000 times to form an approximate chance sampling distribution. The average number of statistically significant findings was 3.23 with a standard deviation of 2.70 (see Fig. 1). Of the 10,000 permutations, only one yielded 24 or more statistically significant findings at the .05 level resulting in a final probability of the set of results of .0001. This number is a *p*-value that can be legitimately interpreted as applying to the *set* of 24 significant correlations, as a group. That is, this *p*-value represents the probability of obtaining a table of behavioral correlates of extraversion such as displayed in Table 1.

The same randomization procedure was employed for the remaining four of the Big Five personality traits and the results for all five are displayed in Table 6. Table 6 reveals several important pieces of information. First, it indicates that the probability of obtaining the observed results for Extraversion, Openness, and Agreeableness are relatively low (below the conventional .05 level). This suggests that we can be reasonably confident that behavior in a 5 min unstructured interaction is related to each of these traits. Additionally, Table 6 indicates that the probabilities of obtaining the observed results for Neuroticism and Conscientiousness are relatively high. This suggests that there is little evidence that behavior in a 5 min unstructured interaction is related to these traits. Further, Table 6 displays the number of statistically significant correlates found on average (Mean Randomly Significant) and the number nominally expected by chance alone. Of note is the similarity between these numbers, suggesting that in fact the number of significant findings one might expect by chance is close to the nominally expected number in these cases. Finally, Table 6 also includes the 95th percentile of the approximate sampling distribution, which is the number of statistically significant findings above which only 5% of the pseudo-samples achieved; this number could be considered a "critical value" of the conventional sort for evaluating the number of significant correlations. However, while setting a critical value is a traditional practice in psychological data analysis, any such threshold is necessarily arbitrary, so we believe that the exact *p*-value is a more informative descriptor of how obtained results compare to those from a chance model.

4. Example #2: Is a particular set of variables related to another particular set of other variables?

While the example provided above demonstrates a solution for research correlating one particular variable with a set of other variables, some designs are more complex. For example, suppose a personality researcher has a host of individual difference measures and is interested in knowing whether they are related, as a group, to a set of outcome variables. Again using the RAP-I data set as a starting point, one might wonder, for example, if personality in general is related to behavior in general. (This is of course the classic "person vs. situation" issue that has long been controversial within personality psychology; see, e.g., Funder, 2008). More specifically, are ratings from the 100 item CAQ predictive of 64 assessments of behavior in an unstructured 5 min interaction? An extension of the randomization method can answer this question. First, each of the 100 CAQ items was correlated with each of the 64 coded behaviors from the RBQ ($N = 163$). Of the resulting 6400 correlations, 608 were statistically significant at the .05 alpha level (two-tailed). The question of interest then is, "What is the probability of getting 608 or more statistically significant results by chance alone?"

The original CAQ profiles provided by the participants were randomly redistributed without replacement such that each participant had an equal probability of being assigned any one of the 163 CAQ profiles to create a pseudo-sample. The pseudo-samples were constructed at the profile rather than the item level so as to keep the profiles of participants consistent. A complete reshuffling

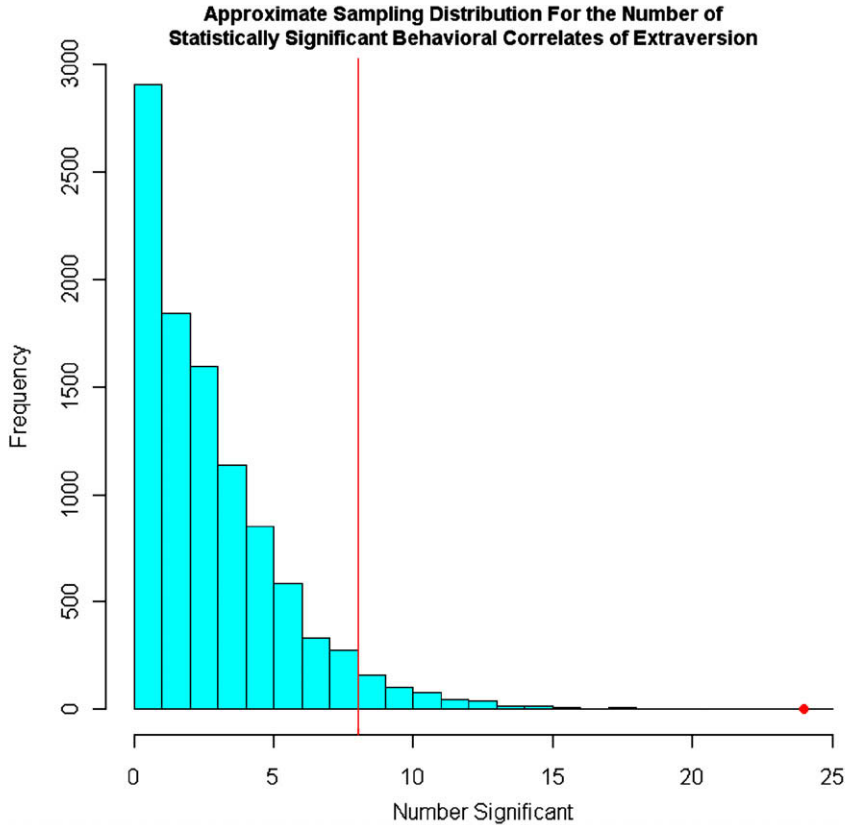

Fig. 1. Based on 10,000 trials. Solid vertical line indicates 95th percentile of distribution. Solid point indicates observed value.

Table 6
Randomization tests for RBQ correlates of the Big Five.

Variable	*E*	*N*	*O*	*C*	*A*
N	159	159	159	159	159
Nominally expected	3.2	3.2	3.2	3.2	3.2
Observed significant (α = .05)	24	1	11	1	12
Mean randomly significant	3.18	3.21	3.21	3.18	3.22
SD	2.67	2.72	2.80	2.67	2.72
p-value	.0001	.8958	.0268	.8977	.0161
95th Percentile	8	8	9	8	8
Trials	10,000	10,000	10,000	10,000	10,000

of the 163 (participants) × 100 (items) matrix is not appropriate because reshuffling in this manner might result in personality profiles that are empirically impossible.[7]

The 6400 correlations from the pseudo-sample were computed and the number of statistically significant correlations was recorded. This procedure was repeated 1000 times to form a distribution of significant findings expected by chance.[8] The average number of statistically significant findings was 319.29 with a standard deviation of 38.73 and a 95th percentile of 387. It is once again interesting to note that the average number of statistically significant findings found within random permutations of the data was very close to the nominally expected number of 320 (.05 × 6400). Of the 1000 permutations, not a single one yielded 608 or more statistically significant findings at the .05 level resulting in a final probability of getting the originally observed 608 or more statistically significant findings of less than 1 in 1000, or $p < .001$. Given the low likelihood of observing 608 or more statistically significant correlations between the CAQ and behavior in a 5 min interaction, it is reasonable to conclude that, in this context, the relationship between personality and behavior is a reliable phenomenon.

5. Example #3: An emphasis on effect sizes

In these first two examples, as well as in Block's (1960) original work on the subject, the key indicator of interest is the number of statistically significant findings at the .05 alpha level. However, the choice of a cutoff for statistical significance is wholly arbitrary and criticisms abound in the literature (e.g. Cohen, 1990, 1994). As an alternative, it might be more reasonable to use an effect size deemed to be of practical or theoretical importance. That is, rather than counting the number of observed statistically significant correlates and comparing the number expected to be significant by chance, one might choose to count the number of observed correlates above an effect size of $r = .10$ (or any other practically or theoretically important value) and compare it to the number of correlates expected to achieve *that effect size* by chance. One advantage of the latter procedure is that significance levels are affected by sample size whereas effect sizes are not, making them perhaps better guides to theory, application, and decisions on where to invest resources in further research.

It is also possible to go further, beyond counting, to raw effect sizes. That is, rather than counting the number of observed significant correlates or counting the number with an effect size greater than some predetermined value, one could compute the mean of the absolute value of the observed effect sizes and compare it to the mean of the absolute value of randomly derived effect sizes.

[7] e.g., a person could ostensibly end up with a score of a one on each item, which is impossible when the Q-sort method is employed (see Block, 2008).

[8] Only 1000 permutations were computed because the computer runtime was already approximately 10 times longer than in the first analysis.

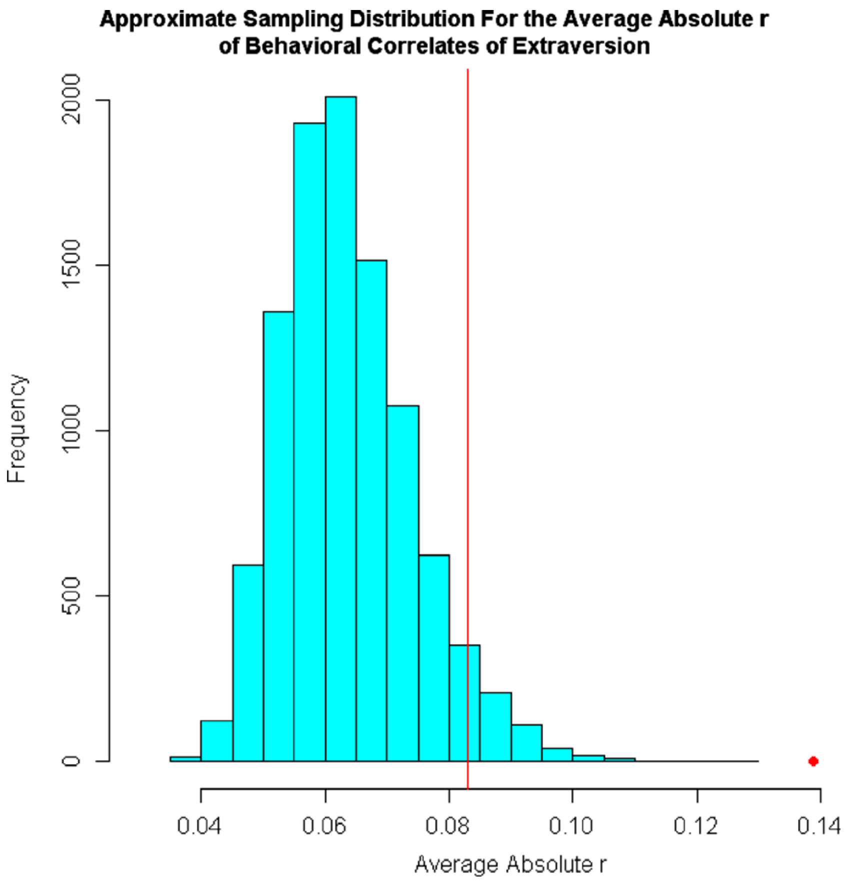

Fig. 2. Based on 10,000 trials. Solid vertical line indicates 95th percentile of distribution. Solid point indicates observed value.

Because this concept may be a bit abstract, a walkthrough of the procedure using Extraversion from Example 1 might be useful.

Is Extraversion related to behavior in a 5 min unstructured interaction? First, the Extraversion scores are correlated with each of the 64 coded behaviors ($N = 159$). Stronger correlations (larger effect sizes) represent better evidence that Extraversion is related to behavior so one might take the average all 64 correlations. However, because both positive and negative behavioral correlates of Extraversion are of interest, one should compute the absolute value of the 64 correlation coefficients before averaging. The resulting number is denoted $|\bar{r}|_{\text{obs}}$. Next, a pseudo-sample is created by randomly redistributing the original Extraversion scores provided by the participants to behavior profiles without replacement such that each profile has an equal probability of being assigned any one of the 159 Extraversion scores and each original score is represented in the simulated data set. The pseudo-sample Extraversion scores are then correlated with each of the 64 behaviors, the absolute values for these correlation coefficients are computed, and the average of the absolute values, denoted $|\bar{r}|_{\text{sim}}$, is computed and recorded. The procedure of random re-assignment, correlation computation, and absolute value averaging and recording is repeated 10,000 times to form an approximate sampling distribution of average absolute values expected by chance (see Fig. 2). The average of this distribution is denoted $\underset{sim}{\bar{\bar{r}}}$ and it is compared to the observed average absolute correlation coefficient ($|\bar{r}|_{\text{obs}}$) to determine the probability of the observed results if the actual relationship between extraversion and behavior was completely random.

We conducted this analysis simultaneously with the Example 1 analysis reported earlier for each of the Big Five traits and the results are displayed in Table 7. Consistent with Table 6, the results of Table 7 indicate that three of the Big Five traits are likely to be related to behavior during a 5 min interaction (Extraversion, Openness, and Agreeableness) while two traits did not show strong evidence of such relationships (Neuroticism and Conscientiousness).

6. Discussion

Widely descriptive investigations that compare many variables to each other often are necessary starting points in research (Funder, 2009). However, researchers, reviewers and readers evaluating such studies are confronted with the difficult task of evaluating the degree to which the results might capitalize on chance. The present paper builds on a proposal by Block nearly 50 years ago to demonstrate how randomization procedures can be extended in such a way as to resolve this dilemma. For example, randomization tests can guide interpretations of correlates of language use as mentioned in the introduction of this paper (Fast & Funder, 2008). In a re-analysis of the data from that study, we found that the personality correlates of the use of certainty words had an overall p-value of .0059 and the behavioral correlates had a p-value of .0002, and the personality correlates of the use of sexuality words had a p-value of .0788 and the behavioral correlates had a p-level of .0025. In this case, the results are reassuring.[9]

6.1. The number of findings expected by chance

The examples presented in this paper extend the work of Block in several ways. First, we use continuous rather than dichotomous data. Additionally, we provide evidence that Block's suggestion

[9] We are grateful to Lisa Fast for sharing the data for this re-analysis.

Table 7
Randomization tests for RBQ correlates of the Big Five using absolute effect sizes.

Variable	*E*	*N*	*O*	*C*	*A*
N	159	159	159	159	159
$\bar{r}_{\text{obs}}$	.1389	.0628	.0895	.0540	.0921
$\bar{\bar{r}}_{\text{sim}}$	.0636	.0636	.0636	.0635	.0636
SD	.0106	.0108	.0109	.0106	.0107
p-value	<.0001	.4717	.0228	.8221	.0139
95th Percentile	.0830	.0832	.0839	.0828	.0832
Trials	10,000	10,000	10,000	10,000	10,000

that the nominal number of findings to be expected by chance may be conservative was, as he cautioned, premature. Each of the analyses conducted here found that the number of significant results expected via the binomial expansion was a close approximation to the average number of significant findings using a randomization test. However, it is important to note that the number of findings to be expected by chance is a property of a given data set and the interdependence of the variables it includes, and must be calculated anew in each research context (a trivially easy task with the use of modern computers). Only further research employing this technique can begin to give us a sense as to whether it is generally true that the nominal and actual numbers of findings to be expected by chance are closely aligned.

6.2. Extending randomization to evaluation of sets of results

A determination of the number of findings expected by chance is just the beginning. Most scientists would wish to establish a more rigorous criterion for accepting results. Reasonable confidence requires estimating a full distribution of chance outcomes and comparing one's results to such a distribution, not just the mean value. In the present data, we found that the 95th percentile of the approximate sampling distributions was consistently 2–3 times greater than the nominal level but this "critical value," again, must be calculated anew for each actual data set. Moreover, we would reiterate that we believe the exact *p*-level of a set of results is more informative than a dichotomous decision as to whether it exceeds an arbitrary threshold.

The present paper extends the use of randomization tests in two other novel ways. Example 2 moves the randomization method from asking, "Is a single variable related to a set of variables?" to "Is a set of variables related to a set of other variables?" Finally, Example 3 demonstrates a randomization approach to evaluating research findings based upon effect sizes rather than arbitrary significance levels.

6.3. "Too many notes"

Researchers interested in broad questions – such as the relationship between personality and behavior – can face ironic difficulties arising from the number of analyses they routinely compute. They may be encouraged by reviewers to reduce their many correlations to a few,[10] or may seek to pre-empt criticism by measuring only a handful of variables (e.g., the Big Five) in the first place. Researchers who wish to explore correlates on a more specific level, such as personality psychologists who employ the 100 items of the CAQ, may confront demands that they perform factor analysis or principal components analysis to reduce the many items and many correlations to a few more general ones. The same advice may be given to researchers interested in numerous behaviors or numerous outcomes in other domains, such as health researchers exploring the connections between aspects of life style and the tendency to contract various diseases and disorders.

While we have nothing against the use of factor analysis per se, we find it odd when its rationale is to reduce data in order to minimize the number of statistics reported. Factors, however derived, are data summaries and like any summary they lose information and blur precision. Any researcher who has sought just the right label for a new factor emerging from her or his data is familiar with these difficulties. Yet the alternative of reporting the full set of correlations in their rich detail risks being disregarded – or unpublished – due to not-unreasonable concerns that they might capitalize on chance. The randomization procedure described here frees researchers of their concerns about measuring too many variables and may lead to deeper investigations that measurement of a few broad traits or factors may miss (e.g. Fast & Funder, 2008).

6.4. Beyond parametric statistics?

We believe that the procedure presented here demonstrates just one out of many potentially valuable uses of randomization tests in psychological research. While debates about whether randomization tests (and other resampling methods) are better or worse than traditional parametric statistics have raged for some time now (e.g. Efron, 1988; Rasmussen, 1988), the clear message is that an understanding of randomization procedures allows researchers to analyze almost any question of interest, particularly when no parametric method exists to do so (Edgington, 1969). We would not be surprised if the coming years see an increasing use of randomization tests over traditional parametric procedures throughout psychology.

Acknowledgments

The authors are also grateful to Jack Block, Daniel Ozer and Lisa Fast for illuminating discussions about the topics discussed here and to Robert Rosenthal, R. Michael Furr, and Jon Grahe for comments on an earlier draft.

All randomization analyses were originally conducted using Statistics101 (version 1.3.3) resampling software which is available for free at http://www.statistics101.net (Grosberg, 2009). The authors are grateful to John Grosberg, creator of Statistics101 software, for his help in developing a more efficient programming code to run the analyses. The JCL used in these analyses is available on request from the first author. A more efficient programming code written in the R language is listed in the Appendix A and is available for download from http://rap.ucr.edu/RandTest.r. All figures were created in the R programming language (R Development Core Team, 2009).

Data from the Riverside Accuracy Project were gathered with the support of NIMH Grant R01-42427. The re- analyses were supported, in part, by NSF Grant 0642243.

[10] The film *Amadeus* includes the following exchange: *Emperor Joseph* II: Your work is ingenious. It's quality work. And there are simply too many notes, that's all. Just cut a few and it will be perfect. *Mozart*: Which few did you have in mind, Majesty?

Appendix A

R Function for Conducting Described Randomization Tests
(*coder comments in italics*)

```
# First enter the two sets of variables in data frame form. Then indicate the number of
# simulations (default 1000) and the critical value of the overall series test (default is 95th
# percentile).

rand.test = function(set1, set2, sims=1000, crit=.95) {

   # This part gets the data ready to be used in the randomization test
   samp.distr=c() #Create a sampling distribution vector for the average absolute r
   samp.distsig=c() #Create a sampling distribution vector for the number statistically significant
   complete = complete.cases(cbind(set1,set2)) # Combine the data sets and keep only complete cases
   set1.set = subset(set1, subset=complete) #Store the “complete” data sets
   set2.set = subset(set2, subset=complete)
   n = nrow(set1.set) #Find the sample size
   critT = qt(.025, n-2, lower.tail=FALSE) # Find the critical t (assumes α = .05) for each test
   critr = sqrt(critT^2 / (critT^2 + n - 2)) # Find the critical r value
   AbsRObs = mean(abs(cor(set1.set, set2.set))) #Find the Avg. Absolute R Obs
   SigObs = sum(abs(cor(set1.set, set2.set)) >= critr) #Find the number significant observed

   # This part starts the randomization
   for (i in 1:sims) {
   rand.order = sample(n, n, replace=FALSE) #Generate a sample of random orders
   cor.mat = cor(set1.set[rand.order,],set2.set) #Get the simulated correlation matrix
   samp.distr[i] = mean(abs(cor.mat)) #Store the absolute average simulated r’s in samp.distr
   samp.distsig[i] = sum(abs(cor.mat) >= critr) #Store the number significant in samp.distsig
   }

   # This part computes the statistical properties of the two sampling distributions
   SimMeanR = mean(samp.distr) #Compute the mean of the sampling distribution
   SimSDr = sd(samp.distr) #And the SD
   Crit95r = quantile(samp.distr,crit) #And the critical value (default 95th percentile)
   pr = sum(samp.distr >= AbsRObs) / sims #Find the probability of the observed value
   SimMeanSig = mean(samp.distsig) #Compute the mean
   SimSDsig = sd(samp.distsig) # SD
   Crit95Sig = quantile(samp.distsig,crit) # Critical value (default 95th percentile)
   pSig = sum(samp.distsig >= SigObs) / sims # Compute a probability value

   #Clean up and print the results
   out.AbsR = round(rbind(n, AbsRObs, SimMeanR, SimSDr, pr, Crit95r),4)
   colnames(out.AbsR) = c(“Average Absolute r”)
   rownames(out.AbsR) = c(“N”, “Observed”, “Exp. By Chance”, “Standard Error”, “p”, “95th%”)
   out.Sig = round(rbind(n, SigObs, SimMeanSig, SimSDsig, pSig, Crit95Sig),4)
   colnames(out.Sig) = c(“Number Significant”)
   rownames(out.Sig) = c(“N”, “Observed”, “Exp. By Chance”, “Standard Error”, “p”, “95th%”)
   print(out.AbsR)
   print(out.Sig)

   # This part creates histogram graphics of the sampling distributions
   old.par = par(mfrow=c(2,1)) # Sets the PAR command two produce two vertical histograms
   hist(samp.distr, freq=TRUE, col=”cyan”, #Create a histogram of the sampling distribution
      main=”Approximate Sampling Distribution /n For Average Absolute r”,
      xlab = “Average Absolute r”, ylab=”Frequency”,
      xlim= range(min(samp.distr)-.01,AbsRObs+.01))
      abline (v=(Crit95r), col=”red”) #Plot the critical value as a line
      points(AbsRObs,0, col=”red”, pch=19) #Plot the observed value point
   hist(samp.distsig, freq=TRUE, col=”cyan”, #Create a histogram of the sampling distribution
      main=”Approximate Sampling Distribution /n For Number Significant”,
   xlab = “Number Statistically Significant”, ylab=”Frequency”,
   xlim= range(min(samp.distsig)-1,(SigObs+1)))
   abline (v=(Crit95Sig), col=”red”) #Plot the critical value as a line
   points(SigObs,0, col=”red”, pch=19) #Plot the observed value point
}
```

References

Blackman, M. C., & Funder, D. C. (1996). Self-esteem as viewed from the outside: A peer and gender perspective. *Journal of Social Behavior and Personality, 11*(1), 115–126.

Block, J. (1960). On the number of significant findings to be expected by chance. *Psychometrika, 25*(4), 369–380.

Block, J. (1961). *The Q-Sort method in personality assessment and psychiatric research*. Springfield, IL: Charles C. Thomas.

Block, J. (1978). *The Q-Sort method in personality assessment and psychiatric research*. Palo Alto, CA: Consulting Psychologists Press.

Block, J. (2008). *The Q-Sort in character appraisal: Encoding subjective impressions of persons quantitatively*. Washington, DC: American Psychological Association.

Brozek, J., & Tiede, K. (1952). Reliable and questionable significance in a series of statistical tests. *Psychological Bulletin, 49*, 339–341.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*(12), 1304–1312.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*(12), 997–1003.

Costa, P. T., & McCrae, R. R. (1985). *The NEO personality inventory manual*. Odessa, FL: Psychological Assessment Resources, Inc..

Cournot, A. A. (1843). *Exposition de la Théorie des Chances et des Probabilitiés*. Paris: Hachette. Reprinted 1984 as Vol. I of Cournot's *Oeuvres Complètes*, ed. B. Bru. Paris: Vrin..

Creed, A. T., & Funder, D. C. (1998a). Social anxiety: from the inside and outside. *Personality and Individual Differences, 25*, 19–33.

Creed, A. T., & Funder, D. C. (1998b). The two faces of private self-consciousness: Self-report, peer-report, and behavioral correlates. *European Journal of Personality, 12*, 411–431.

Eaton, L. G., & Funder, D. C. (2003). The creation and consequences of the social world: An interactional analysis of extraversion. *European Journal of Personality, 17*, 375–395.

Eaton, L. G., & Funder, D. C. (2001). Emotional experience in daily life: Valence, variability, and rate of change. *Emotion, 1*(4), 413–421.

Edgington, E. S. (1969). *Statistical inference: The distribution free approach*. New York: McGraw-Hill.

Efron, B. (1988). Bootstrap confidence intervals: Good or bad? *Psychological Bulletin, 104*(2), 293–296.

Fast, L. A., & Funder, D. C. (2008). Personality as manifest in word use: Correlations with self-report, acquaintance report, and behavior. *Journal of Personality and Social Psychology, 94*(2), 334–346.

Fast, L. A., Funder, D.C. (in press). Gender differences in the correlates of self-referent word use: Authority, entitlement, and depressive symptoms. *Journal of Personality*.

Fisher, R. A. (1935). *The design of experiments*. Oxford, England: Oliver & Boyd.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102*, 652–670.

Funder, D. C. (2008). Persons, situations and person-situation interactions. In O. P. John, R. Robins, & L. Pervin (Eds.), *Handbook of personality* (3rd ed., pp. 568–580). New York: Guilford.

Funder, D. C. (2009). Persons, behaviors, and situations: An agenda for personality psychology in the postwar era. *Journal of Research in Personality, 43*(2), 120–126.

Funder, D.C. (in press). Naive and obvious questions. *Perspectives on Psychological Science*.

Funder, D. C., Furr, R. M., & Colvin, C. R. (2000). The riverside behavioral Q-sort: A tool for the description of social behavior. *Journal of Personality, 68*(3), 451–489.

Funder, D. C., Kolar, D. C., & Blackman, M. C. (1995). Agreement among judges of personality: Interpersonal relations, similarity, and acquaintanceship. *Journal of Personality and Social Psychology, 69*(4), 656–672.

Furr, R. M., & Funder, D. C. (1998). A multimodal analysis of personal negativity. *Journal of Personality and Social Psychology, 74*, 1580–1591.

Furr, R. M., & Funder, D. C. (2004). Situational similarity and behavioral consistency: Subjective, objective, variable-centered and person-centered approaches. *Journal of Research in Personality, 38*, 421–447.

Grosberg, J. (2009). Statistics101 (version 1.3.3) [computer software, <www.statistics101.net>].

Markey, P. M., Funder, D. C., & Ozer, D. J. (2003). Complementarity of interpersonal behaviors in dyadic interactions. *Personality and Social Psychology Bulletin, 29*, 1082–1090.

Miller, R. G. (1966). *Simultaneous statistical inference*. New York: Wiley.

R Development Core Team, (2009). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Rasmussen, J. L. (1988). Bootstrap confidence intervals: Good or bad: comment on efron (1988) and strube (1988) and further evaluation. *Psychological Bulletin, 104*(2), 297–299.

Rodgers, J. L. (1999). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. *Multivariate Behavioral Research, 34*(4), 441–456.

Rosenthal, R., & Rubin, D. B. (1984). Multiple contrasts and ordered Bonferroni procedures. *Journal of Educational Psychology, 76*, 1028–1034.

Sakoda, J. M., Cohen, B. H., & Beall, G. (1954). Test of significance for a series of statistical tests. *Psychological Bulletin, 51*, 172–175.

Schimmack, U., Oishi, S., Furr, R. M., & Funder, D. C. (2004). Personality and life satisfaction: A facet level analysis. *Personality and Social Psychology Bulletin, 30*, 1062–1075.

Spain, J. S., Eaton, L. G., & Funder, D. C. (2000). Perspectives on personality: The relative accuracy of self versus others for the prediction of emotion and behavior. *Journal of Personality, 68*, 837–867.

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. New York: John Wiley & Sons Inc..

Wilkinson, B. (1951). A statistical consideration in psychological research. *Psychological Bulletin, 48*, 156–158.